Multi Experiment Matrix - web tool for mining co-expressed genes over hundreds of datasets

Priit Adler[a], Raivo Kolde[bc], Meelis Kull[bc], Hedi Peterson[ac], Jüri Reimand[b], Aleksandr Tkatšenko[bc] and Jaak Vilo[bc]

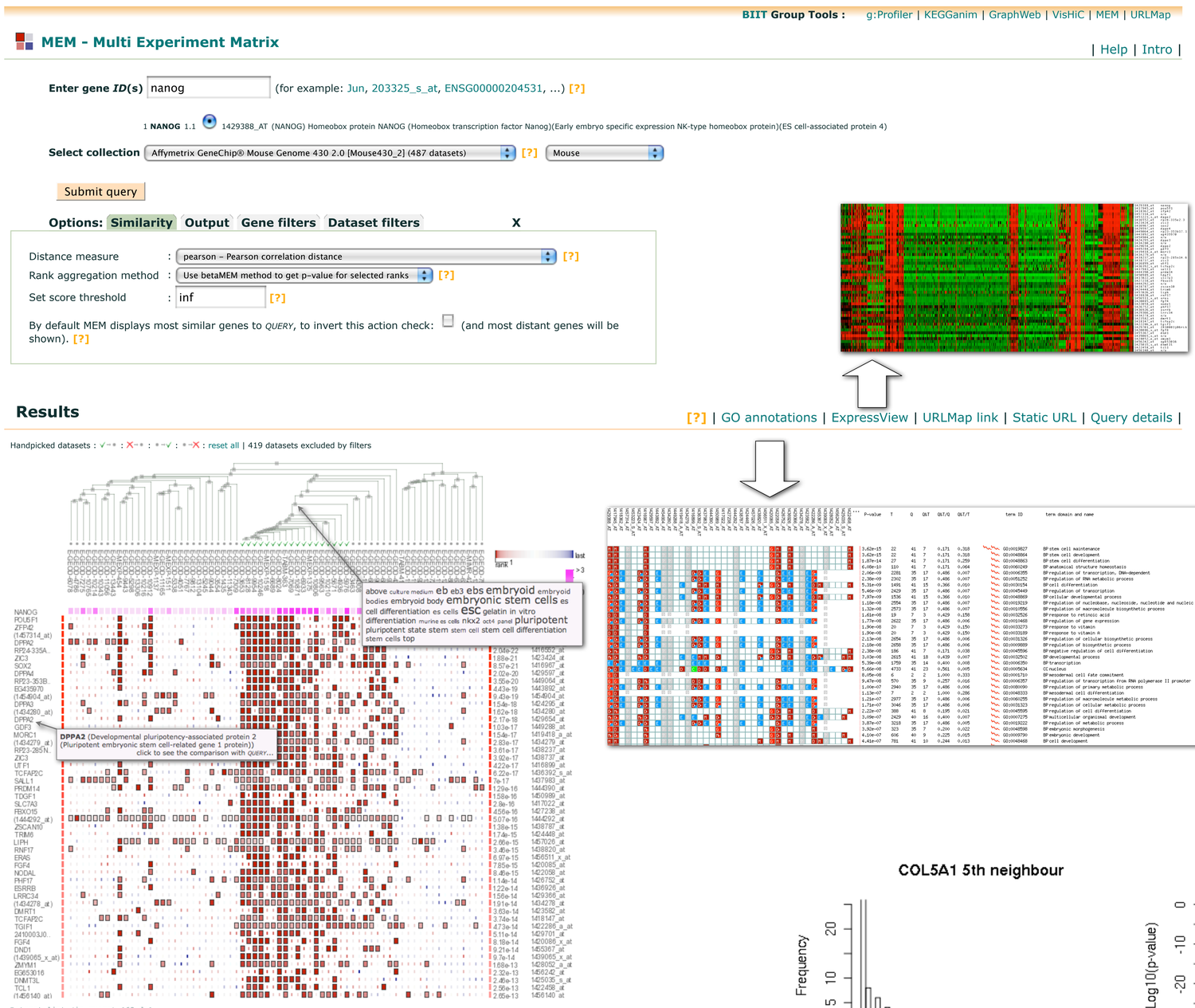
[a] Institute of Molecular and Cell Biology, University of Tartu,
[b] Institute of Computer Science, University of Tartu,
[c] Quretec Inc

biit.cs.ut.ee/mem

Abstract

Accumulation of gene expression data in public domain has raised the opportunity to discover new facts by re-analyzing existing experiments. Co-expression over many gene expression datasets has been proven useful in many areas of molecular biology and bioinformatics, such as network reconstruction and gene function prediction.

MEM query goes over hundreds of gene expression datasets from ArrayExpress covering wide spectrum of biological conditions from stem cells to different diseases and cancer. The essence of MEM is a novel rank aggregation method that combines similarity searches across individual data sets into a global ordering by selecting automatically the appropriate experiments and assigning a combined p-value for the significance of the similarity across all data.

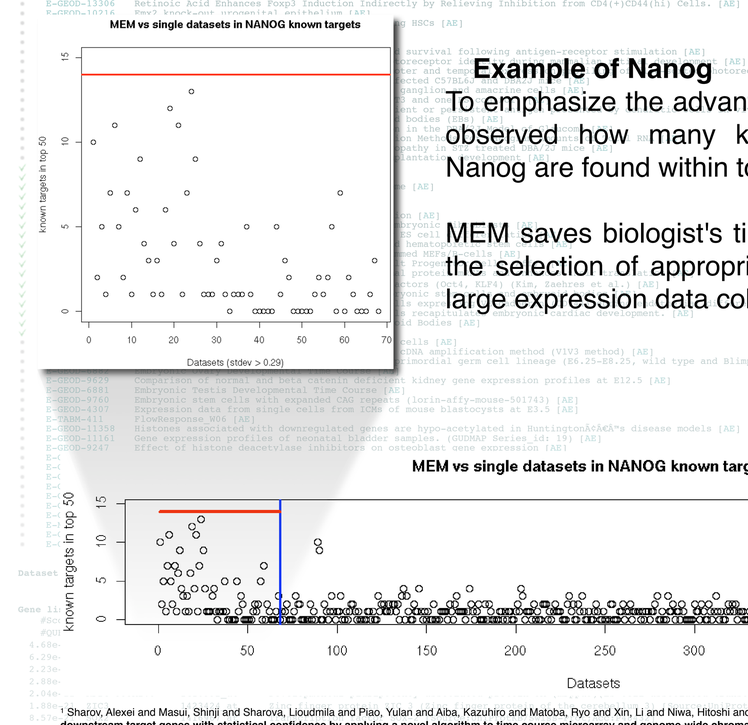


Method

Query gene by user
In each dataset genes are sorted according to a query gene
- ✓ The distance measure used in this step can be arbitrary

The results are converted into ranks
- ✓ Overall we get a rank matrix and for each gene we end up a rank vector over all datasets
- ✓ We expect two different distributions from rank vector:
  - uniform if there is no connection between genes (e.g. Collagen's 6250th neighbour)
  - skewed towards 0 if genes are co-expressed (e.g. Collagen's 5th neighbour)

Ranks are aggregated into p-values
- ✓ With the p-value measures deviation from uniform distribution. More small ranks leads to smaller p-value.



Example of Nanog

To emphasize the advantages of MEM, we observed how many known targets[1] of Nanog are found within top fifty.

MEM saves biologist's time by automating the selection of appropriate datasets from large expression data collections.

Example of Collagen

To illustrate the rank aggregation method behind MEM query we used *Collagen alpha-1(V) chain Precursor*(COL5A1) as an example. The rank distribution and dynamics of p-value are depicted on the graph for fifth and 6250th result in MEM query. For each gene we have a rank vector over all datasets.
- ✓ Ranks are sorted and for each rank a p-value is calculated to measure how skewed towards the 0 it is.
- ✓ Minima of the p-values is presented as MEM score.

[1] Sharov, Alexei and Masui, Shinji and Sharova, Lioudmila and Piao, Yulan and Aiba, Kazuhiro and Matoba, Ryo and Xin, Li and Niwa, Hitoshi and Ko, Minoru. **Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data.** *BMC Genomics,* vol 9, 2008

Data

From ArrayExpress database we have downloaded all AffyMetrix GeneChip® gene expression datasets with raw data available.
- ✶ Datasets are grouped by platform which makes features of different experiments easy to compare
- ✶ Many popular platforms have hundreds of different experiments with wide spectrum of conditions

Conclusion

We have developed a tool for finding co-expressed genes over many public gene expression datasets. The tool has many possible applications, such as gene function prediction or network reconstruction. The search is powered by a novel rank aggregation algorithm. User interface is highly interactive and provides additional information about the query and the results.

Acknowledgements

This work has been supported by EU FP6 ENFIN, FunGenES, Estonian Science Foundation 5724, 7437

Priit Adler has been awarded a Travel Fellowship by ISMB/ECCB 2009 to present this poster.